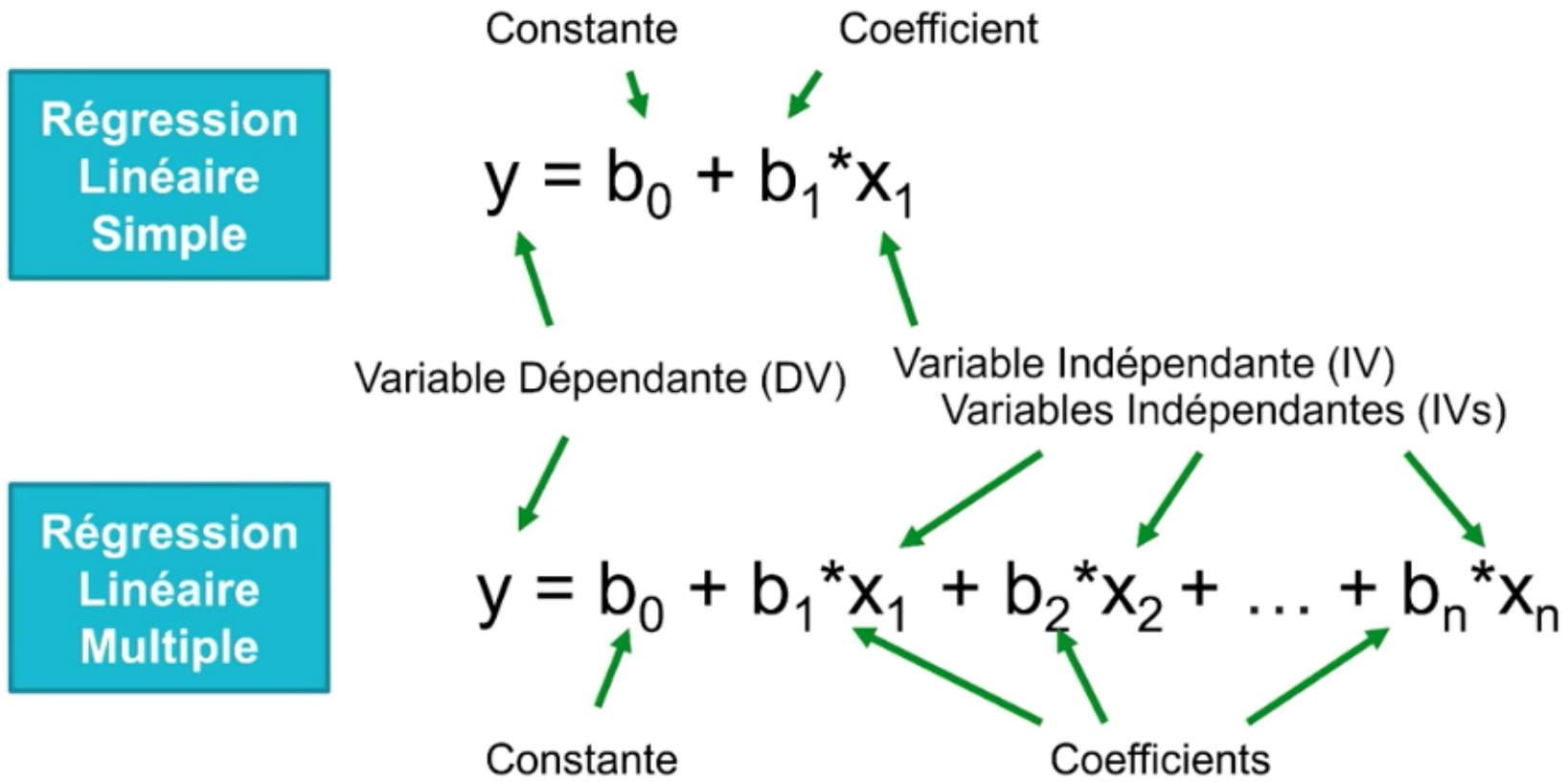
Machine Learning and Data Science

Régression linéaire multiple

Bassem Ben Hamed

Juillet 2018

Régression Linéaire



Attention

Hypothèses de la Régression Linéaire:

1. Exogénéité
2. Homoscédasticité
3. Erreurs indépendantes
4. Normalité des erreurs
5. Non colinéarité des variables indépendantes

Les Dummy Variables

Profit	Dépenses R&D	Admin	Marketing	Etat
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	Californie
191,050.39	153,441.51	101,145.55	407,934.54	Californie
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	Californie

Dummy Variables

New York	Californie
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1$$

Le piège des Dummy Variables

Profit	Dépenses R&D	Admin	Marketing	Etat
192,261.83	165,349.20	136,897.80	471,784.10	New York
191,792.06	162,597.70	151,377.59	443,898.53	Californie
191,050.39	153,441.51	101,145.55	407,934.54	Californie
182,901.99	144,372.41	118,671.85	383,199.62	New York
166,187.94	142,107.34	91,391.77	366,168.42	Californie

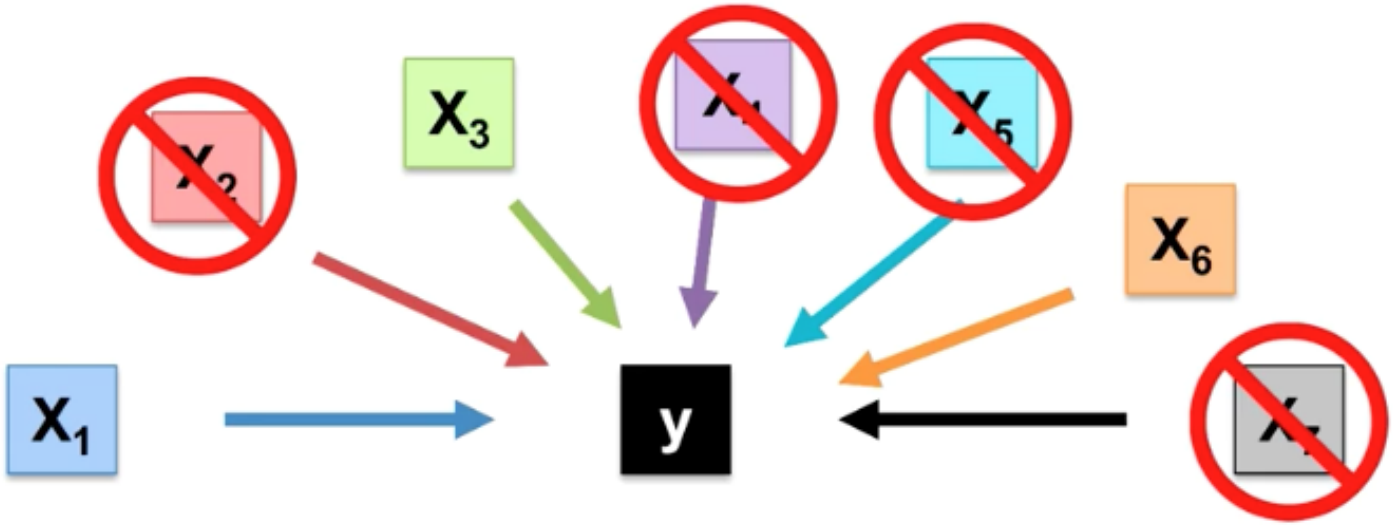
Dummy Variables

New York	Californie
1	0
0	1
0	1
1	0
0	1

$$y = b_0 + b_1 * x_1 + b_2 * x_2 + b_3 * x_3 + b_4 * D_1 + b_5 * D_2$$

Toujours enlever une dummy variable

Construire Un Modèle



Pourquoi ?

Construire un Modèle

5 méthodes de construction de modèles:

1. All-in
2. Backward Elimination
3. Forward Selection
4. Bidirectional Elimination
5. Score Comparison

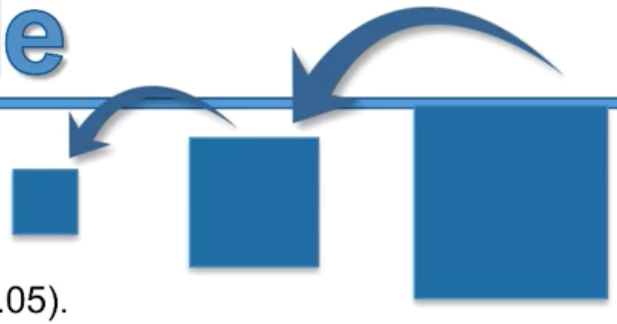
} Stepwise Regression

Construire Un Modèle

“All-in”

- Vous savez déjà ce qu'il faut mettre
- Vous n'avez pas le choix
- Vous voulez vous préparer pour la Backward Elimination

Construire Un Modèle



Backward Elimination

STEP 1: Choisir un seuil SL pour rester dans le modèle (e.g. SL = 0.05).

STEP 2: Remplir le modèle de tous les prédicteurs possibles

STEP 3: Considérer le prédicteur ayant la plus grande p-value
Si p-value > SL, aller au STEP 4, sinon c'est FINI

STEP 4: Enlever le prédicteur

STEP 5: Ajuster le modèle sans cette variable

FIN: Votre modèle est prêt

Construire Un Modèle

Forward Selection

STEP 1: Choisir un seuil pour entrer dans le modèle (e.g. SL = 0.05)

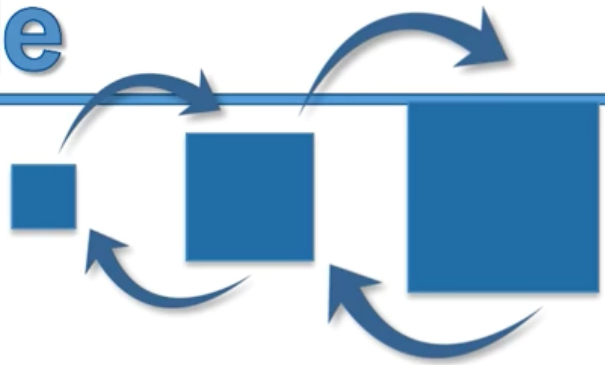
STEP 2: Ajuster tous les modèles simples de regression $\mathbf{y \sim x_n}$
Sélectionner celui avec la plus petite p-value

STEP 3: Garder cette variable et ajuster tous les modèles possibles avec un prédicteur en plus

STEP 4: Considérer le prédicteur ayant la plus petite p-value
Si p < SL, aller au STEP 3, sinon c'est FINI

FIN: Garder le modèle précédent

Construire Un Modèle



Bidirectional Elimination

STEP 1: Choisir deux seuils pour entrer (Ex: SLENTER = 0.05) et rester (SLSTAY = 0.05) dans le modèle

STEP 2: Effectuer le next step de la Forward Selection (les nouvelles variables doivent vérifier: p < SLENTER pour entrer dans le modèle)

STEP 3: Effectuer TOUS les steps de la Backward Elimination (les vieilles variables doivent vérifier p < SLSTAY pour rester dans le modèle)

STEP 4: Aucune nouvelle variable peut entrer et aucune ancienne variable peut sortir

FIN: Votre modèle est prêt

Construire Un Modèle

All Possible

STEP 1: Choisir un critère de qualité d'ajustement (ex: critère d'Akaike)

STEP 2: Construire tous les modèles de régression possibles: 2^N-1 combinaisons au total

STEP 3: Choisir celui ayant le meilleur critère

FIN: Votre modèle est prêt

Exemple:
10 colonnes donnent
1023 modèles